Study on Post-processing Method for HMM-based Phonetic Segmentation using Adaptive Neuro Fuzzy Inference system

By Liang Dong

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Professor Masato Akagi

September, 2014

Study on Post-processing Method for HMM-based Phonetic Segmentation using Adaptive Neuro Fuzzy Inference system

By Liang Dong (1210215)

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Professor Masato Akagi

and approved by
Professor Masato Akagi
Professor Jianwu Dang
Associate Professor Masashi Unoki

August, 2014 (Submitted)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The contemporary speech technology and research heavily depend on large speech corpora, which were segmented and labelled at phonetic level. In general, a speech database consist of speech recordings and some transcriptions, including word transcription, phonetic transcription and phone level time-alignment. In contrast to other kind of transcriptions, phonetic time-alignment is considered as a difficult task and also plays an important role [1]. In speech technology, such as automatic speech recognition (ASR) and text to speech (TTS) synthesis, phonetic time-alignment is used for the training of acoustic models. While in linguistics, phonetic time-alignment is used to do research into phonetic, conversation analysis, dialectology and other fields [2]. The requirements of speech segmentation differ on different applications and researches. But it is believed that good segmentation can provide significant benefits to both speech technology and research.

Actually, highly accurate and reliable speech segmentation is much more needed for most applications. Although in some kind of application, such as speech recognition, highly accurate of segmentation seems not urgent needed. Because HMM training is an averaging process that tends to smooth segmentation errors [3]. However, in concatenative speech synthesis, an essential approach to TTS synthesis, a segmentation error may produce an audible error in the synthetic voice. The final quality of synthesized speech heavily depends on the accuracy of speech segmentation.

Traditionally, manual segmentation has been considered the most reliable and precise method to get the segments for speech corpus. In addition, it is used as the standard for the evaluation of automatic speech segmentation. However, manual segmentation is time-consuming and labor-intensive that can be performed only by expert phoneticians [4]. Otherwise, compared with automatic speech segmentation, it is not stable and the results of segmentation influenced by environmental and human factors. Especially, in the age of big data, the shortcoming of manual segmentation is a fatal defect for large speech corpus. There is a strong demand for us to develop automatic speech segmentation.

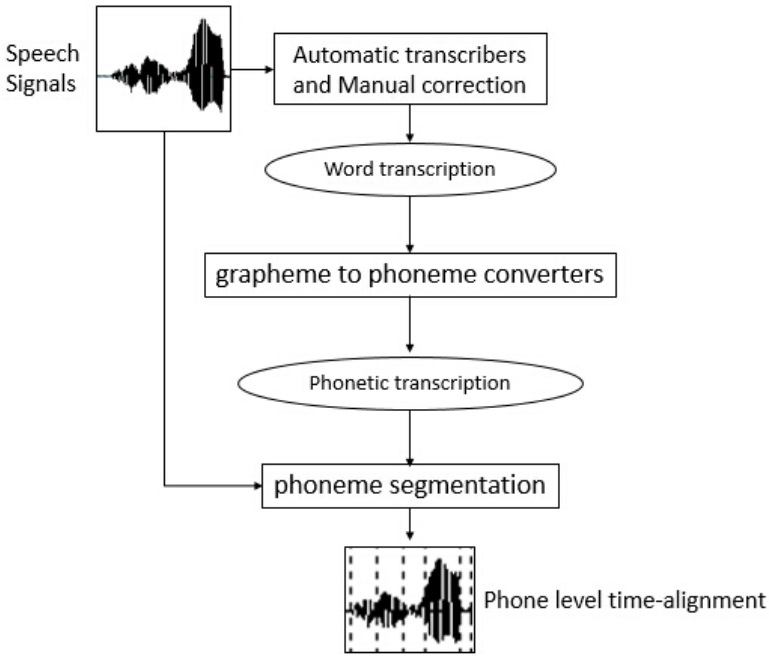


Figure 1.1: The basic work flow of speech segmentation.

1.2 Background

Phoneme segmentation is the ability of human beings to break words down into individual sounds. It is a kind of phonological awareness skills, not only essential for humans, but also an important technique for speech research and technology. It can be applied in many areas, such as: automatic speech segmentation (ASR), Text-to-speech (TTS) synthesis, speech corpus segmentation and labelling.

The basic work flow of speech segmentation as shown in Fig. 1.1. Firstly, word transcriptions are extracted from the speech signal by automatic transcribers and some manual corrections. Secondary, phonetic transcriptions can be obtained by grapheme to phoneme converters from word transcriptions. Finally, phoneme segmentation produces the phone-level time-alignment. Different from word transcriptions and phoneme transcriptions, phoneme segmentation is always considered as a difficult task. The reason is that the boundary between phonemes may exist at some processing level in our brain, but not in the acoustic signal [5].

In general, there are two kinds of speech segmentation methods: manual speech segmentation and automatic speech segmentation. The comparison of these two methods as shown in table 1.1. Manual speech segmentation is always considered as the most accurate method, and often used as the evaluation criteria for automatic speech segmentation. But it is not only time-consuming and labor-intensive, but also not a stable method. To improve the accuracy of automatic speech segmentation is much more needed.

Different kinds of approaches have been proposed for the task of automatic speech

Table 1.1: The comparison between manual speech segmentation and automatic speech segmentation.

Method	Manual speech segmentation	Automatic speech segmentation
Advantage	Most reliable and accurate.	Fast and stable.
Disadvantage	Time-consuming and labor-intensive, not stable.	Not accurate.

segmentation, such as: the detection of variations/similarities in spectral [6] or prosodic parameters of speech [7], the template matching using dynamic programming or the synthetic speech [8] and the discriminative learning segmentation [9].

Nowadays, the mainstream technology for automatic speech segmentation is called forced alignment that needs the recorded audio and phone or word sequence as input. As seen in Fig. 1.1. in this method, Hidden Markov Model (HMM) is commonly used to build a model for each phoneme with different number of states [10]. Speech signal is exacted as a set of feature vectors by frame. The alignment of frames with phonemes is determined by finding the most likely sequence of hidden states given the observed data and the acoustic model represented by the HMMs. The model can give the rough boundaries for each phoneme. The reported performance of traditional HMM-based forced alignment systems range from 80% to 89% agreement within 20ms compared to manual segmentation on the TIMIT corpus [11] [22].

Actually, HMM-based forced alignment produces a segmentation which seems to be precise enough to train the HMMs for speech recognition, because HMM training is an averaging process that tends to smooth segmentation errors [3]. However, in speech synthesis a segmentation error may produce an audible error in the synthetic voice, that has led speech synthesis to rely on manual segmentation for years [28]. The drawback of HMM is that phone boundaries are not represented in the model for the phonetic segmentation of the HMM-based forced alignment [15].

Many methods have been proposed to improve accuracy of automatic speech segmentation within HMM framework, by refining the initial HMM-based segmentation[3-5]. Some research has achieved great improvement for traditional HMM-based segmentation [16]. Some statistical correction procedures were used in this research, for the reason that the criteria for TIMIT boundary assignments stated that [25]:" *The boundary between many semivowels and their adjacent vowels is ratherill-defined in the waveform and spectrogram, because transitions are slow and continuous. It is not possible to define a single point in time that separates the vowel from the semivowel. In such case we decided to adopt a simple heuristic rule, in which one-third of the vocalic region is assigned to the semivowel.*"

Actually, some researchers were aware that the difference between HMM-based segmentation and expert phoneticians is the rules and knowledge for segmentation. It is as same as the heuristic rules used in TIMIT database, which can be the key point for the improvement of automatic speech segmentation.

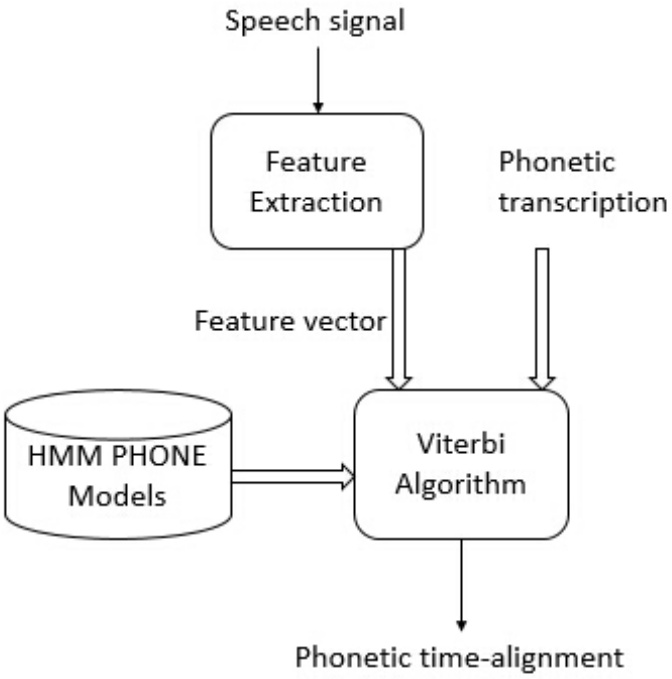


Figure 1.2: The HMM-based forced alignment.

Fuzzy logic provides a very straightforward way to introduce human knowledge(which is fuzzy by nature) in computer-based systems [17]. So they simulated the process of manual speech segmentation by manually defining the fuzzy rules [18]. For example, they use the fuzzy rules as:

$$If\ ENERGY\_RIGHT\ is\ LOW\ then\ TRANS\_PROB\ is\ HIGH$$

$$If\ ENERGY\_LEFT\ is\ HIGH\ then\ TRANS\_PROB\ is\ HIGH$$

to represent the well known knowledge used by human that a voiced-unvoiced transition is a transition from a high energy region to a low energy region. The results show that it is a good way to improve the automatic segmentation. But the drawback is that the fuzzy rules need to be carefully designed and adjusted by experts [12].

1.3 Definition of problems

1.3.1 Human segmentation of speech

Even phoneticians, the nearest to the ideal in speech segmentation and labelling, need to identify the phonemes and determine the position of a phonetic boundary by examine different features. In fact, humans could not determine the time mark of phonetic boundary in some kind of situation, such as the boundary between semivowels and their adjacent. Some transitions between phonemes are continuous and slow, in such case, prior experience and knowledge are needed for the segmentation.

What is more, humans divided the task of segmentation into two different steps [18]:

- Step 1: Listening to the speech in order to identify the phoneme sequence and roughly determine the position of a phonetic boundary.
- Step 2: Visually examining several features of the speech signal (waveform, energy, spectrogram, etc.) around the rough position of the phonetic boundary determined in step 1 in order to place the phonetic boundary time mark where these features best satisfy certain conditions specific for that kind of phonetic boundary.

By listening, phoneticians only can determine the phoneme sequence with the rough boundaries. It means the most important knowledge for the precise segmentation exists in the second step. Both the examination of some special acoustic features and the experience for the determination of unclear boundaries are needed.

1.3.2 HMM-based speech segmentation

In HMM based speech segmentation, it is assumed that the sequence of observed speech vectors corresponding to each phoneme is generated by a Markov model as shown in Fig. 1.2. This figure shown a five states hidden Markov model, the observation sequence extracted from the speech signal by frame. The Mel-frequency cepstral coefficients (MFCCs) are commonly used for the extraction. For each HMM, there is a state transition to describe the connection between the HMM and a corresponding output. For each phoneme, a HMM with a given number of states is used, including the non-emitting entry and exit states. The parameters for each state in the HMM can be trained by Baum-Welch algorithm [24].

Figure 1.3 summarises the use of HMMs for phonetic segmentation. A HMM is trained for each phoneme using a number of examples of that phoneme. Then, to recognise the time mark of a sequence of phonemes, the likelihood of each model generating that phoneme is calculated and the most likely model identifies the phoneme.

1.3.3 Drawbacks of HMM-based segmentation

There has been brief description for human speech segmentation and HMM-based segmentation method in the last two subsections. The drawbacks of HMM-based segmentation are stated in this subsection.

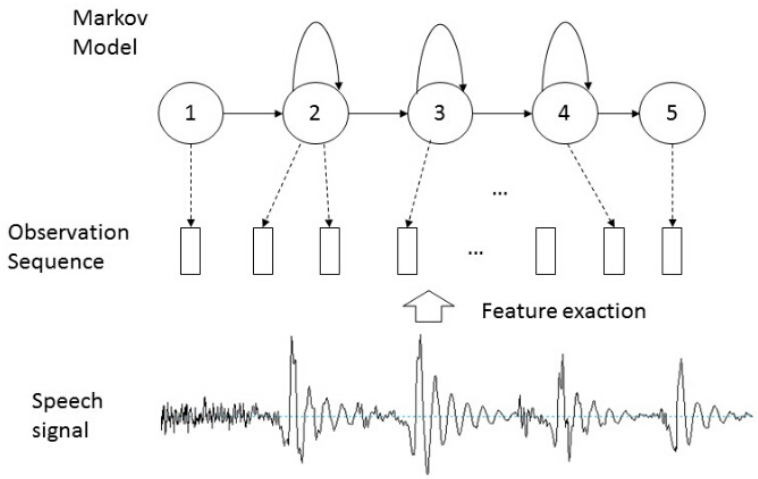


Figure 1.3: The Markov generation model for a phoneme.

As seen in the last section, although the HMM performs well in the representation of speech. However, it is designed for the recognition, there is no clearly definition for the boundary of phoneme in this model. In addition, HMM training is an averaging process that tends to smooth segmentation errors which ensure it could give us a good performance in segmentation. But the averaging process also loss the diversity of phoneme segmentation.

Compared with manual segmentation, HMM-based segmentation only can give us a rough boundary between two phonemes through probability calculation. For the reason that the trained HMMs for each phoneme is an average model for the voice with different gender, age and accents. It could not give more precise time mark of boundaries. Further more, the human knowledge and experience for segmentation, especially the examining different features around the phoneme boundary is so important for the precise segmentation. Based on its drawbacks, an appropriate post-processing method is needed for the improve of automatic segmentation within HMM framework.

1.4 Purpose of this research

The results by HMM-based automatic segmentation method are often found unsatisfactory to be directly applied to TTS or some other applications with the requirement for highly accurate and reliable speech segmentation. Thus, the boundaries produced by traditional HMM segmentation should be refined by some post-processing method, in order to obtain highly accurate boundaries. According to the Subsection 1.3.2, feature examining and prior experience exists in the manual speech segmentation, which ensure the accurate of segmentation. If this mechanism can be modelled, it will be applicable to obtain the boundaries approximated to manual speech segmentation, to solve the problem

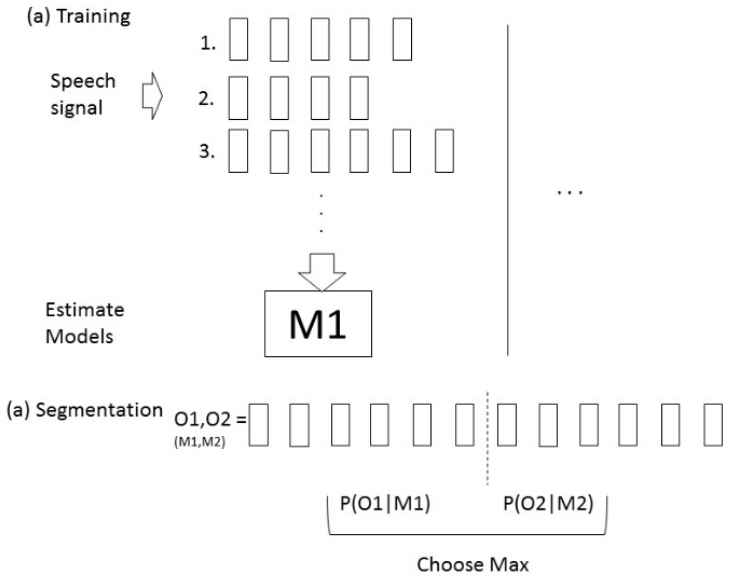


Figure 1.4: Using of HMMs for phonetic segmentation.

that segmentation based on HMM.

Fuzzy Inference Systems (FIS) are efficient techniques for studying the behavior of nonlinear systems using fuzzy logic rules. Fuzzy logic provides a very straightforward way to introduce human knowledge(which is fuzzy by nature) in computer-based systems [17]. Adaptive neuro fuzzy inference system (ANFIS) is a Neuro-fuzzy system that uses the learning techniques of neural networks, with the efficiency of fuzzy inference systems [26]. Compared with other learning methods, it has both the advantages of neural networks and fuzzy inference systems, and also have a better performance than others [27]. Some research simulate the process of manual speech segmentation by manually define the fuzzy rules [18]. Although it has some improvement, the fuzzy rules need to be carefully designed and adjusted by experts [19]. Compared with it, ANFIS can be trained quickly and in a completely automatic way. Based on its advantages: simple implementation and good performance in learning nonlinear and fuzzy rules, it is very suitable to solve our problems.

In this study, the main purpose is to improve the accuracy of traditional HMM-based speech segmentation. we used adaptive neuro fuzzy inference system to compensate the arbitrariness in manual speech segmentation and the systematic segmentation errors produced by HMMs. To learn such kind of rules between manually labeled and forced alignment boundaries to improve the traditional HMM-based automatic phoneme segmentation, some acoustic features are selected to train the system. In order to compare the final results with the outstanding works, we use the same condition of the research [16] in our experiment. Context-independent(monophone) phone models were used to train the HMM. Well trained adaptive neuro fuzzy inference system is used to refine the time

marks of HMM-based forced alignment in order to obtain a higher accuracy of the final result for our proposed method. Further more, the influence of different features used for training, including PLP and MFCC, were also investigated. Meanwhile, the different frame rate were used for the extraction of features, in order to obtain higher accuracy.

1.5 Thesis structure

The rest of the thesis is organised as follows:

- Chapter 2: An outline of Adaptive Neuro Fuzzy Inference System will be described in this chapter.
- Chapter 3: To achieve the goal presented in the last section, an outline of the proposed method is introduced.
- Chapter 4: This chapter introduces the database used in the research. Moreover, the different between original 61 phonemes set and 54 phonemes set are described. The mapping mechanism of phoneme set is introduced in the subsection of 3.2.
- Chapter 5: How to extract phoneme boundaries and implement labelling are described in this chapter. The results of experiments for segmentation, using TIMIT database, are also shown in this chapter.
- Chapter 6: Based on the results of evaluation, the advantages and disadvantages for the proposed method are summarized in this chapter.
- Chapter 7: In the last chapter, the conclusion for this thesis is presented. Moreover, according to the current work, future work is put forward to.

Chapter 2

Adaptive Neuro Fuzzy Inference System

According to Section 1.4, the Adaptive Neuro Fuzzy Inference System (ANFIS) has both advantages of Neuro network and Fuzzy Inference system, will be adopted for HMM-based automatic speech segmentation in our study. Although the architecture and learning procedure of adaptive networks are well described in [29], this chapter will introduce an outline of the model.

2.1 Basic concept

The typical structure of ANFIS as shown in Fig. 2.1. In this structure the inputs are x_1 and x_2, y is the only one output. Each node in the same layer has the same function. The $O_{1,i}$ represents the output of layer 1 and i-th node. The details of each layer as shown below.

- Layer 1: This layer let the input signal be fuzzy.

$$O_{1,i} = u_{A_i}(x_1), i = 1, 2 \tag{1}$$

$$O_{1,j} = u_{B_{j-2}}(x_2), j = 3, 4 \tag{2}$$

  A_i and B_{j-2} are the fuzzy sets, represent for the concept such as 'many' or 'less'. The $u_{A_i}(x_1)$ is the membership function of fuzzy set. Trapezoidal, triangular and bell-shaped functions are commonly used membership functions.

- Layer 2: This layer calculate the firing strength of a rule, which multiplies the incoming signals and sends the product out.

$$O_{2,i} = w_i = u_{A_i}(x_1))u_{B_i}(x_2), i = 1, 2 \tag{3}$$

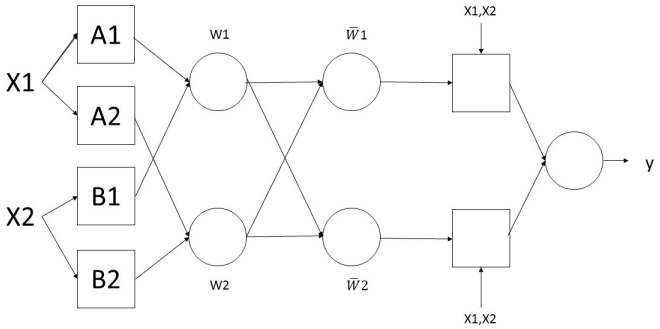


Figure 2.1: The structure of ANFIS.

- Layer 3: This layer is the normalized calculation of all rules firing strengths. The i-th node calculates the ratio of the i-th rule?s firing strength to the sum of all rules firing strengths.

$$O_{3,i} = \bar{w_i} = w_i/(w_i + w_i), i = 1, 2 \tag{4}$$

- Layer 4: Every node in this layer is a square node with a node function to compute the output of each rule.

$$O_{4,i} = \bar{w_i} f_i = w_i(p_i x_i + q_i x_2 + r_i), i = 1, 2 \tag{5}$$

- Layer 5: The single node in this layer is a circle node, that computes the overall output as the summation of all incoming signals.

$$O_{5,i} = y = \sum \bar{w_i} f_i, i = 1, 2 \tag{6}$$

Compared with fuzzy logic made by experts, it is a method based on data. The fuzzy rules are generated from data, not by inaccurate experience and intuition. It is fit for the system that characteristics do not fully understand by human or too complex to define, such as the speech segmentation in our study.

Chapter 3

The proposed method

HMM-based automatic segmentation method are often found not accurate enough to be directly applied to some applications. By comparing the different between manual segmentation and traditional HMM-based speech segmentation, we found the post-processing is the key point for the highly accurate. The purpose of this research is to propose an automatic speech segmentation method of highly accuracy. Based on the advantage of ANFIS, we proposed a post-processing method for HMM-based segmentation. The outline of the proposed method is constructed in following section.

3.1 Outline of the proposed method

In manual speech segmentation, heuristic rules are used to deal with unclear boundaries. Moreover, HMM-based forced alignment always has systematic errors which need to be corrected. In our method, ANFIS is used to compensate for these errors by refining HMM-based forced-alignment.

A schematic description of our segmentation and labelling system is shown in Fig. 3.1. It is divided into two steps: Firstly, the system obtains the initial time marks from the HMM-based forced-alignment. Secondly, it refines the time marks by a well trained ANFIS.

3.2 First step: HMM-based forced alignment

The aim of the first step is to get the initial time marks. HMM-based forced alignment was built by HTK toolkit as our baseline system [24]. For constructing a speech segmentation system, acoustic features are an important factor. Therefore, the most relevant acoustic feature PLP which have been successful in related works were selected [15]. 39 PLP were used for training features including 13-dim PLP and the first and second derivatives were extracted. Following [16], the number of state for each phoneme as shown in Table 3.1. Eight Gaussian mixtures per state were used. Then, state-tied Monophone GMM-HMMs were trained based on maximum likelihood(ML). For the reason that our system was trained on utterances, we only compared the results based on utterances not individual

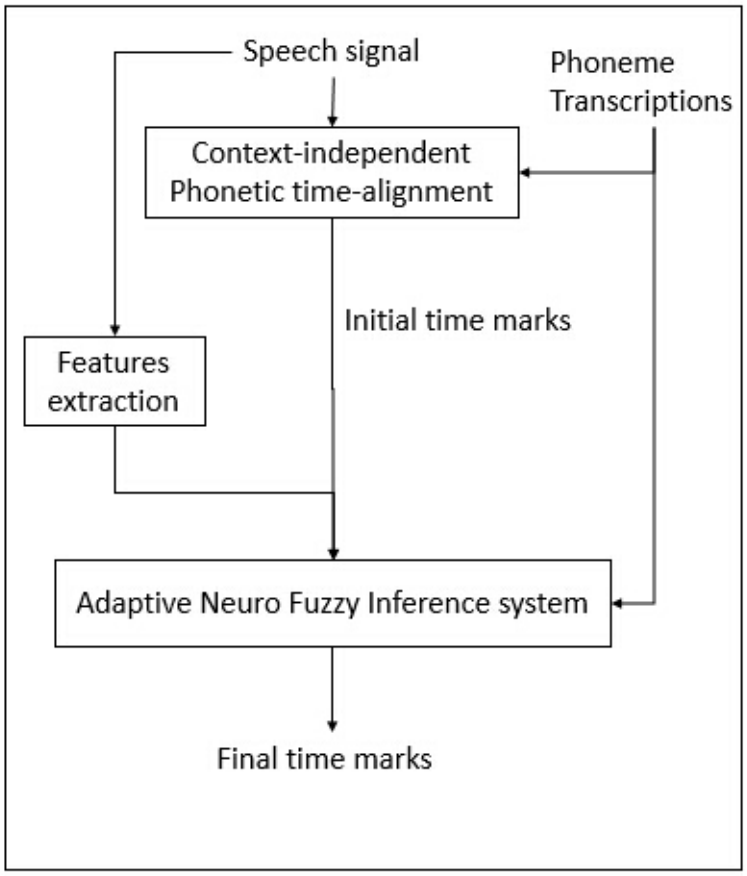


Figure 3.1: The segmentation and labelling system.

phones in [15]. After training, the HMMs were used to obtain the initial time marks for both training data and testing data.

3.3 Second step: ANFIS-based refinement

In the second step, we need to solve two problems. As [25] said that separating the vowel from the semivowel at a point in time is very difficult. In such case they adopted a simple heuristic rule, in which one-third of the vocalic region is assigned to the semivowel. To compensate for such arbitrariness, a linear model was built to correct the forced alignment boundaries between vowel/glide phonemes in one research [15]. The model predicts manual boundary positions from the forced alignment positions of the two phonemes (phoneme center positions), the identities of the boundaries (the phonemes preceding and following the boundary), and the forced alignment boundary positions. Another problem is the systematic segmentation errors produced by HMMs which was solved by the mean difference between manually labeled and forced alignment boundaries in [15].

To compensate the arbitrariness in manual speech segmentation and to correct the

Table 3.1: The HMM-based forced alignment.

Phone models:	Monophone HMMs.
1-state HMMs:	/pcl/ /bcl/ /tcl/ /dcl/ /kcl/ /gcl/ /axh/ /l/ /r/;
5-state HMMs:	/ay/ /aw/ /oy/;
3-state HMMs:	others.

systematic segmentation errors produced by HMMs, we used an ANFIS in our proposed method. In our method, ANFIS is used to train the difference between HMM forced alignment and manual segmentation, in order to estimate the refining boundary. Compared with other learning methods, it has both the advantages of neural networks and fuzzy inference systems, and always has a better performance than others [27].

The Architecture of ANFIS is presented in [29]. Sugeno's fuzzy if-then rules [30] are used. A typical fuzzy rule in a Sugeno fuzzy model has the format

$$If x \text{ is } A \text{ and } y \text{ is } B \text{ then } z = f(x, y),$$

where A and B are fuzzy sets in the antecedent; z = f(x,y) is a crisp function in the consequent.

The training of ANFIS is very important and often affect the final result. Effective input and output help ANFIS to learn the useful knowledge in speech segmentation. Manual speech segmentation is clearly a multi-input single-output process. Similarly, in our method, Multi-input and single-output ANFIS is built for each phoneme in the 54 phonemes set. Each ANFIS has its Sugeno fuzzy model based on its inputs and output. Besides, MATLAB tools [31] are used to train the ANFIS.

In manual segmentation, some features contains important information helps the human of the determination for the final decision. The information of distance and speech signal are obviously used in manual segmentation. So, two kinds of features are used in our training: distance features and acoustic features. Distance features are computed based solely on the forced alignment time marks while acoustic features are computed based solely on the speech signal.

The distance features give us more information about the classification of different situations. For distance features, we consider the phoneme on the left of a boundary(L_P) and the phoneme on the right of a boundary(R_P). The duration of L_P and R_P, the phoneme type of R_P and the position by forced alignment(t) are used. Also, we use the number from 0 to 53 to represent the 54 different phonemes. These kinds of information are used to divided the time marks of forced alignment into different types.

It is well know that different phonemes have different acoustic features, by comparing acoustic features before and after a time mark give us the information about the phoneme boundaries. In this way, the log energy and the first two PLP are selected, which are the most important features to represent a phoneme [18]. We will use C0, C1 and C2 to represent these three features in short. Acoustic features at a time position are computed based

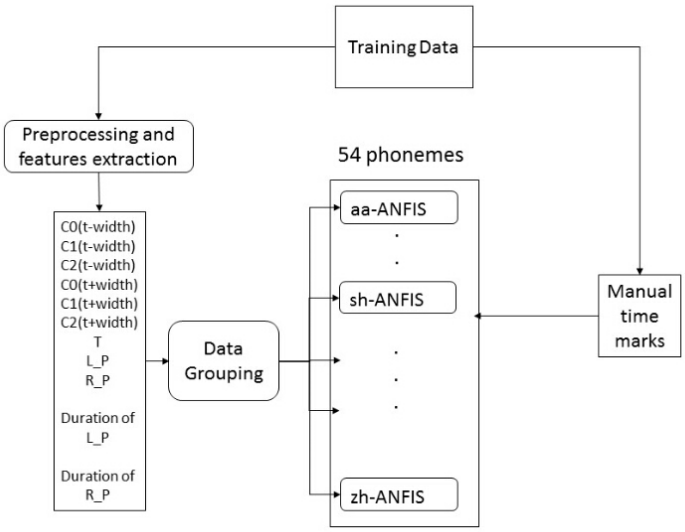


Figure 3.2: The input and output with ANFIS architecture.

on two windows of fixed width, one to the right(t+width) and one to the left(t-width) of that time position. The parameter for the width will be discussed in experiments. The features including C0, C1 and C2 which are computing by RASTA-PLP [23]. Totally, we have six inputs of signal features: C0(t-width), C1(t-width), C2(t-width), C0(t+width), C1(t+width) and C2(t+width). Only the manual time marks were selected as the output. Before training, all of the inputs and outputs were normalized between 0 and 1. Then, each phoneme was trained individually. An ANFIS architecture for one of the 54 phonemes with ten inputs and one output as shown in Fig. 3.2. After training, the system was used to refine the forced alignment boundaries by HMM-based forced alignment.

Chapter 4

Database

There will be a brief introduction for the databased used for this research in this chapter. TIMIT database is commonly used for the research of automatic segmentation which was used in our experiments. Besides, for the reason that unlike the 61 phonemes set used in TIMIT database, many researchers use 54 phonemes set in their studies. Two kinds of phoneme set classification were compared in our experiments. The mapping strategy from 61 phonemes set to 54 phoneme set will be introduced in the following section.

4.1 TIMIT database

Table 4.1: The phoneme set (61 phonemes).

Type	Phoneme
Pauses and stop closures	pau, pcl, bcl, tcl, dcl, kcl, gcl, h#, eqi, q
Vowels	aa, ae, ah, ao, aw, ax, ax-h, axr, ay, eh er, ey, ih, ix, iy, ow, oy, uh, uw, ux
Glides	l, r, w, y, hh, hv, el
Nasals	m, n, ng, nx, em, en, eng
Plosives	b, d, g, p, t, k, dx, jh, ch
Fricatives	s, z, sh, zh, f, v, th, dh

The TIMIT corpus of read speech is designed to provide speech data for acoustic phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16kHz speech waveform file for each utterance. The speech was recorded at Texas Instruments, Inc (TI), transcribed at Massachusetts Institute of Technology (MIT) [21].

Transcriptions of TIMIT corpus are made by audiologists and have been hand verified. In this research, the phonetic transcriptions will be treated as the most precise labelling

results and used for comparison to automatic segmentation results. Excluding the "dialect calibration" sentences, test subset (including 1340 sentences) and training subset (including 3696 sentences), balanced for phonetic and dialectal coverage, are specified [21]. In this experiments, all test sentences will be used.

4.2 Phoneme mapping

The complete set of 61 TIMIT phoneme symbols (as shown in table 4.1) was mapped to a set of 54 phonemes as follow [22]. First, the sentence-beginning and sentence-ending pause were mapped to pause (/pau/). Epenthetic silence (/epi/) was also mapped to pause. The syllabic phonemes /em/, /en/, /eng/, and /el/ were mapped to their non-syllabic counterparts /m/, /n/, /ng/, and /l/, respectively. The glottal closure symbol /q/ was removed. If the glottal closure neighbored a voiced phoneme on one side and an unvoiced phoneme on the other side, the glottal closure was merged with the voiced phoneme. If the glottal closure was surrounded by two voiced phonemes, then the boundary of the two neighboring phonemes was placed at the mid-point of where the glottal closure occurred. If the glottal closure was surrounded by two unvoiced phonemes, it was mapped to a short neutral vowel, /ax/. Finally, short pauses with duration less than 20 msec were removed. If the short pause neighbored a voiced phoneme on one side and an unvoiced phoneme on the other side, the short pause was merged with the unvoiced phoneme. Otherwise, the boundary of the two neighboring phonemes was placed at the mid-point of where the short pause occurred. The algorithm of phoneme mapping as shown in Fig. 4.1.

4.3 Evaluation

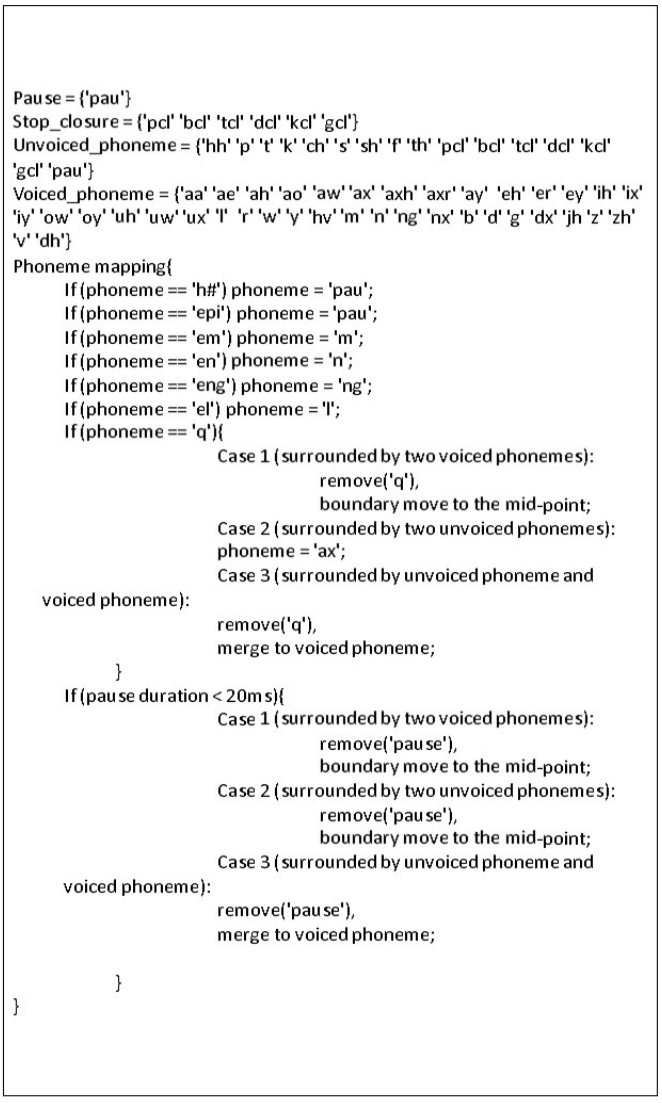
In this study, the manual segmentation is considered as the criterion for the evaluation of automatic speech segmentation. The accuracy rate is used for the comparison, which is calculated by the formula:

$$Accuracyrate = Number\ of\ correct\ boundaries\ /Total\ number\ of\ boundaries$$

An estimated boundary is correct when

$$|Estimated\ Position\ -\ Manual\ Position| < Tolerance$$

Then, the accuracy rate of boundaries can be calculated statistically based on the results.



```
Pause = {'pau'}
Stop_closure = {'pcl' 'bcl' 'tcl' 'dcl' 'kcl' 'gcl'}
Unvoiced_phoneme = {'hh' 'p' 't' 'k' 'ch' 's' 'sh' 'f' 'th' 'pcl' 'bcl' 'tcl' 'dcl' 'kcl'
'gcl' 'pau'}
Voiced_phoneme = {'aa' 'ae' 'ah' 'ao' 'aw' 'ax' 'axh' 'axr' 'ay'  'eh' 'er' 'ey' 'ih' 'ix'
'iy' 'ow' 'oy' 'uh' 'uw' 'ux' 'l'  'r' 'w' 'y' 'hv' 'm' 'n' 'ng' 'nx' 'b' 'd' 'g' 'dx' 'jh' 'z' 'zh'
'v' 'dh'}
Phoneme mapping{
        If (phoneme == 'h#') phoneme = 'pau';
        If (phoneme == 'epi') phoneme = 'pau';
        If (phoneme == 'em') phoneme = 'm';
        If (phoneme == 'en') phoneme = 'n';
        If (phoneme == 'eng') phoneme = 'ng';
        If (phoneme == 'el') phoneme = 'l';
        If (phoneme == 'q'){
                        Case 1 (surrounded by two voiced phonemes):
                                remove('q'),
                                boundary move to the mid-point;
                        Case 2 (surrounded by two unvoiced phonemes):
                        phoneme = 'ax';
                        Case 3 (surrounded by unvoiced phoneme and
    voiced phoneme):
                        remove('q'),
                        merge to voiced phoneme;
                }
        If (pause duration < 20ms){
                        Case 1 (surrounded by two voiced phonemes):
                                remove('pause'),
                                boundary move to the mid-point;
                        Case 2 (surrounded by two unvoiced phonemes):
                                remove('pause'),
                                boundary move to the mid-point;
                        Case 3 (surrounded by unvoiced phoneme and
        voiced phoneme):
                        remove('pause'),
                        merge to voiced phoneme;

                }
}
```

Figure 4.1: The algorithm used for the phoneme mapping.

Chapter 5

Experiments for segmentation

The experiments were divided into two parts for our proposed method. In the first experiment, the HMM-based forced alignment was built in order to obtain the initial time marks for both training and testing data. In addition, two kinds of acoustic features including MFCCs and PLP were used. The second experiment including training the ANFIS and refine the initial time marks of testing data. The aim is to improve the accuracy of the HMM-based forced alignment in the first experiment. The details will be described in the following sections.

5.1 The baseline system

The aim of baseline system is to construct a traditional HMM system to obtain initial time marks which has a higher accuracy. The experiments including the data preprocessing, parameters optimization and the discussion to determine the final baseline system.

5.1.1 Preprocessing for using TIMIT database

When using TIMIT database, some preprocessing should be implemented.

1. The TIMIT speech waveform files are SPHERE-headed [20] and can not be used in HTK and directly. Thus, firstly the waveform files should be SPHERE-striped to obtain the raw matrices and rewritten by the function "wavwrite" in MATLAB. Meanwhile, the speech waveform was digitized to 16 bits at a sampling frequency of 8KHz.

2. The start time and the end time for each phoneme in the phonetic transcriptions are recorded at the sampling points. They are unable to be used in HTK, because the unit of time is 100 nanoseconds in HTK [24]. Besides, they are not suitable for comparison to the results by the proposed method, because the unit of time is 1 millisecond in the results by the proposed method. Thus, there should be a small change for the original phonetic transcriptions.

3. As in subsection of 3.2 described, we mapped the 61 phonemes set into 54 phonemes set for TIMIT database. The 54 phonemes set as show in Table 5.1. Besides, excluding

Table 5.1: The 54 phonemes set.

Pauses and stop closures	/pau/	/pcl/	/bcl/	/tcl/	/dcl/	/kcl/	/gcl/
Vowels	/aa/	/ae/	/ah/	/ao/	/aw/	/ax/	/axh/
	/axr/	/ay/	/eh/	/er/	/ey/	/ih/	/ix/
	/iy/	/ow/	/oy/	/uh/	/uw/	/ux/	
Glides	/l/	/r/	/w/	/y/	/hh/	/hv/	
Nasals	/m/	/n/	/ng/	/nx/			
Plosives	/b/	/d/	/g/	/p/	/t/	/k/	/dx/
	/jh/	/ch/					
Fricatives	/s/	/z/	/sh/	/zh/	/f/	/v/	/th/
	/dh/						

the "dialect calibration" sentences, 3,696 utterances were used for training and 1344 utterances were used for testing.

5.1.2 Description for experiments

The actual training process takes place in stages and it is illustrated in more detail in Fig. 5.1.

1. Firstly, an initial set of models must be created, an example as shown in Fig. 5.2.

2. The extraction of MFCCs and PLP were described in the later section. The training data of MFCCs or PLP was used as bootstrap data.

3. The tools HInit and HRest provide isolated word style training (context-dependent HMM training) using the fully labelled bootstrap data. Each of the required HMMs is generated individually. HInit reads in all of the bootstrap training data and cuts out all of the examples of the required phone. It then iteratively computes an initial set of parameter values using a segmental k-means procedure. The tool HCompV can be used for a so-called flat start, obtain the global speech mean and variance. In this case, all of the phone models are initialised and have state means and variances equal to the global speech mean and variance.

4. HVite uses the token passing algorithm to perform Viterbi-based speech recognition. It use phoneme sequences as an input, a dictionary defining the phonemes set with HMMs. Then, it attaching the appropriate HMM definition to each phone instance. By computing, the results of forced alignment were obtained.

5.1.3 Feature extraction

For constructing a speech segmentation system, acoustic feature is an important factor needed to be investigated. Therefore, the most relevant acoustic features Melfrequency cepstral coefficients (MFCCs), and PLP features using RASTA processing (RASTA-PLP) which have been successful in related works were selected [23]. Besides, the frame rate

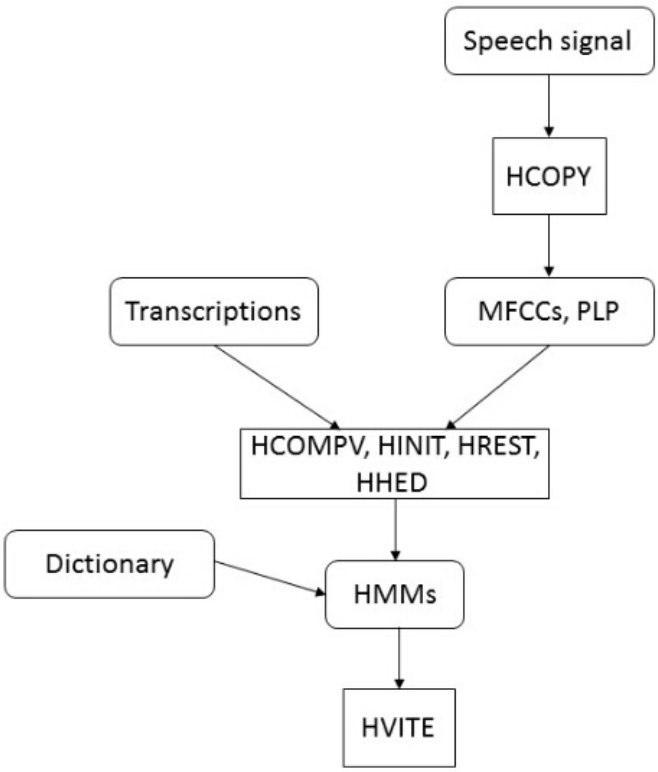


Figure 5.1: The training process.

used to extract acoustic features were also discussed in our experiment, in order to obtain higher accuracy before the refinement by ANFIS.

Frame rate

This experiment was constructed to investigate the relationship between frame rate used for extracting features and the performance of trained HMMs. The aim is to optimize the parameter of frame rate. Monophone HMM and GMM acoustic models were used as described in section 2.2. The standard 39 MFCC were extracted with 25ms Hamming window, which is commonly used in the training of HMMs. The 1ms, 10ms and 20ms were used as the frame rate for comparison.

The results as shown in Fig. 5.3. The blue line indicates the accuracy rate of HMMs, which were trained by the 10ms frame rate MFCC. Compared with the 1ms frame rate and 20ms frame rate, it has a better performance in the training of HMMs. Finally, we choose 10ms as the parameter for the frame rate. This parameter was used in the following experiments.

```
~o <VecSize> 39 <MFCC_0_D_A>
~h "aa"
<BeginHMM>
<NumStates> 5
<State> 2
<NumMixes> 8
<Mixture> 1 0.125
<Mean> 39
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
<Variance> 39
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
<Mixture> 2 0.125
<Mean> 39
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
<Variance> 39
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
```

Figure 5.2: A part of prototype model for phoneme aa.

PLP and MFCC

Whether use PLP or MFCC for the training of HMM need to be considered. The experiment was construct to compare these two kinds of acoustic features. Both PLP and MFCC were used for the training of HMMs. The parameters used to extract the standard 39 MFCC or PLP features as shown in table 5.2.

The results of accuracy rate, including MFCC and PLP show the agreement between forced alignment and manually labeled boundaries are listed in Table 5.3.

Compared to the model trained by MFCC and PLP, PLP is a little better than MFCC for 20ms tolerance. However, poor performance for 5ms, 10ms and 15ms tolerance. Actually, the difference between these two are small enough that can be ignored. Finally, we choose PLP to construct our proposed method.

5.1.4 Conclusion

After some discussions in the previous experiments, the baseline system used to obtain the initial time marks was determined. The Monophone HMM and GMM acoustic models, with the standard 39 PLP features extracted with 25ms Hamming window and 10ms frame rate, were trained using the HTK toolkit. The accuracy rate as shown in table.

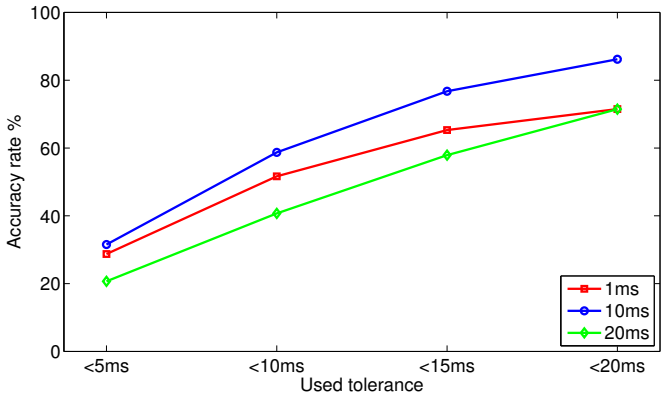


Figure 5.3: The different duration used for the extraction of acoustic features.

Table 5.2: The parameters for the extraction of MFCCs and PLP.

	MFCC	PLP
TARGETKIND	MFCC_0_D_A	PLP_0_D_A
TARGETRATE	100000	100000
WINDOWSIZE	250000	250000
USEHAMMING	T	T
PREEMCOEF	0.97	0.97
NUMCHANS	20	20
CEPLIFTER	22	22
NUMCEPS	12	12
USEPOWER	-	T
LPCORDER	-	12

5.2 Refinement by ANFIS

Humans refine the rough boundaries by comparing the features before and after the rough time marks between two phonemes. So the most important work for this section is to find the most suitable features that used for training. We assume that if the features below to a same phoneme have something in common, otherwise it will not. By optimizing, 10ms was used to select the features before and after the rough boundaries got from the standard HMMs.

5.2.1 Data preparation

In this experiment, we got the HMM-based labeling files from the trained HMM speech segmentation model with the manual labeling files. The model that use PLP with 25 Hamming window and 10ms frame rate at the previous experiment was used. Phoneme type, log energy, PLP features were used as input for training the Adaptive Neuro Fuzzy

Table 5.3: Agreement percentage for different tolerance of MFCC and PLP.

	<5ms	<10ms	<15ms	<20ms
MFCC	31.53%	58.71%	76.76%	86.18%
PLP	31.37%	58.39%	76.63%	86.25%

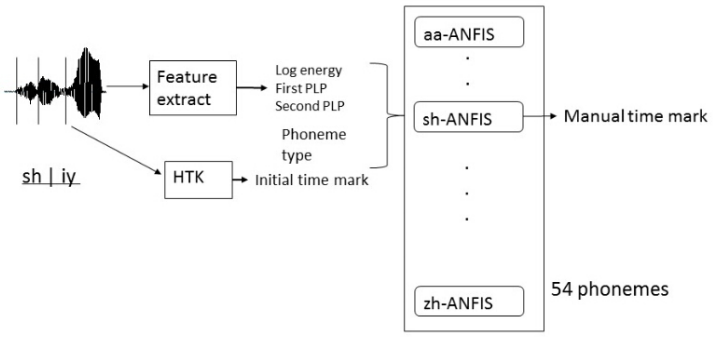


Figure 5.4: An example for training ANFIS.

Inference System. The manual labeled time marks is used as output. After training, we use the system to refine the time marks got from previous experiment.

5.2.2 Features extraction

An example for training the ANFIS as show in Fig. 5.4. The boundaries between sh and iy were used for the training of sh-ANFIS, which is the ANFIS for the phoneme sh. The phonemes type, acoustic features and some distance information were used as inputs for training. The manual time marks is used as the only one output.

We assume that the acoustic features before and after the initial time marks could give us the information for the refinement. But the duration before and after the initial time marks should be determined for the extraction of features. So, we use 5ms, 10ms, 15ms and 20ms as duration to extract acoustic features for training. The results as shown in table 5.4. We can see the duration of 10ms has a higher accuracy rate in all tolerance. In this case, 10ms was used as the final parameter for the duration.

5.2.3 ANFIS training

The ANFIS can be trained for several times. For the first training, we can use the initial time marks with other features as inputs and manual time marks as output. After that, the refined time marks can be used as the input for the second training. And the manual time marks for boundaries are always used as the output for training.

In this experiment, we trained the ANFIS twice. For the reason that there is no much

Table 5.4: Different duration used for the extraction of acoustic features

	<5ms	<10ms	<15ms	<20ms
5ms	46.46%	73.76%	85.81%	91.59%
10ms	47.08%	74.02%	86.01%	91.60%
15ms	46.52%	73.16%	85.71%	91.59%
20ms	43.89%	71.39%	84.98%	91.43%

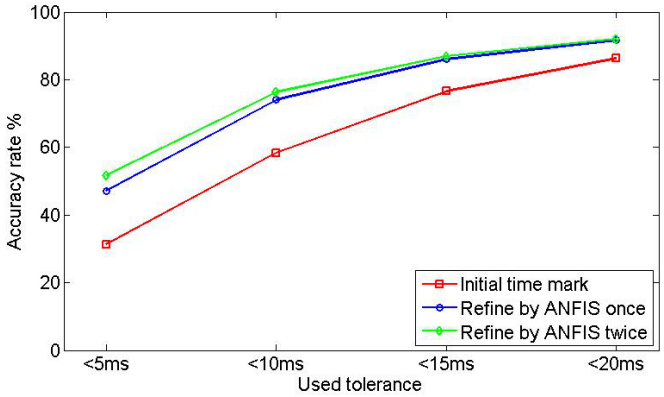


Figure 5.5: The refinements by ANFIS.

improvement of the accuracy rate in 20ms tolerance. We did not train it more. The results as shown in Fig. 5.5. We can see the second refinement by ANFIS, the improvement within 20ms tolerance is a little. But there is still some improvement in the small tolerance. It means if we need higher accuracy within a small tolerance, use ANFIS repeatedly can be effective.

5.2.4 Results

Following the proposed method mentioned in the subsection 3.3, the refined time marks of HMM-based forced alignment were obtained. By comparing with manual segmentation, the accuracy rate as shown in table 5.4. Besides, the accuracy rate for different kinds of boundaries as shown from the Fig. 5.6 to Fig. 5.11.

Table 5.5: Comparing the accuracy of the traditional HMM and the combination of HMM and ANFIS.

Tolerances	<5ms	<10ms	<15ms	<20ms
Traditional HMM	31.37%	58.39%	76.63%	86.25%
The proposed ANFIS	51.71%	76.32%	86.93%	92.08%

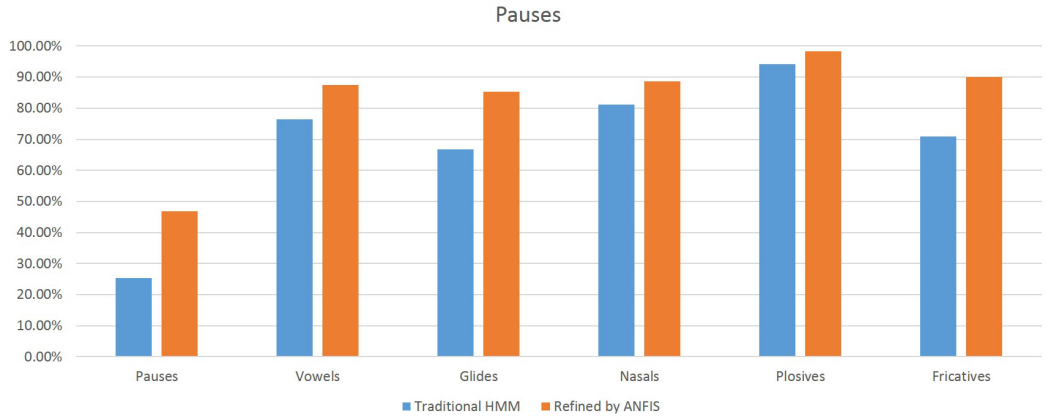


Figure 5.6: The accuracy of boundaries between Pauses and others within 20ms tolerance.

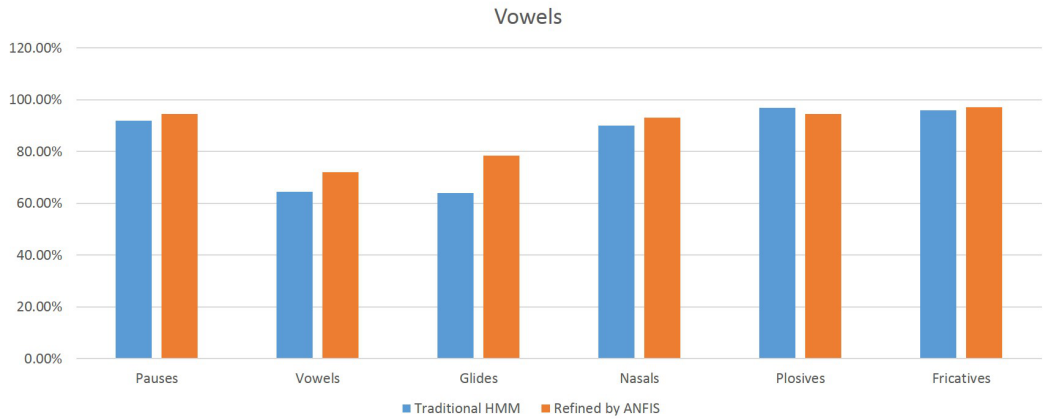


Figure 5.7: The accuracy of boundaries between Vowels and others within 20ms tolerance.

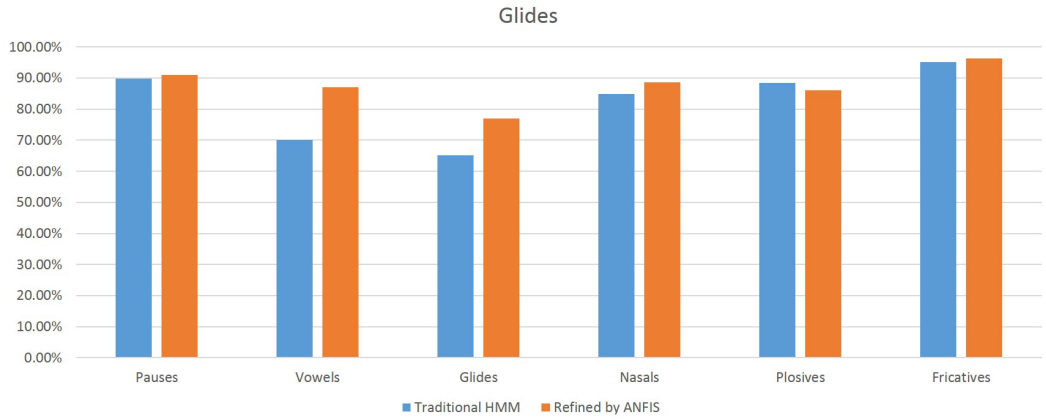


Figure 5.8: The accuracy of boundaries between Glides and others within 20ms tolerance.

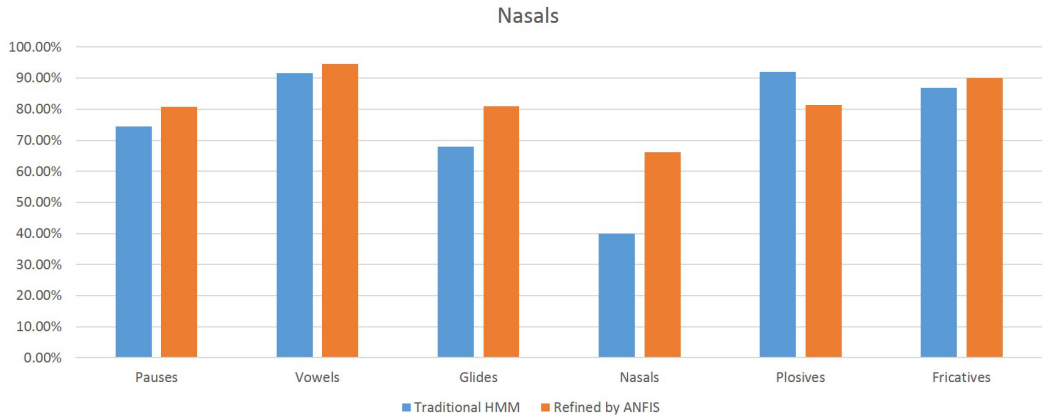


Figure 5.9: The accuracy of boundaries between Nasals and others within 20ms tolerance.

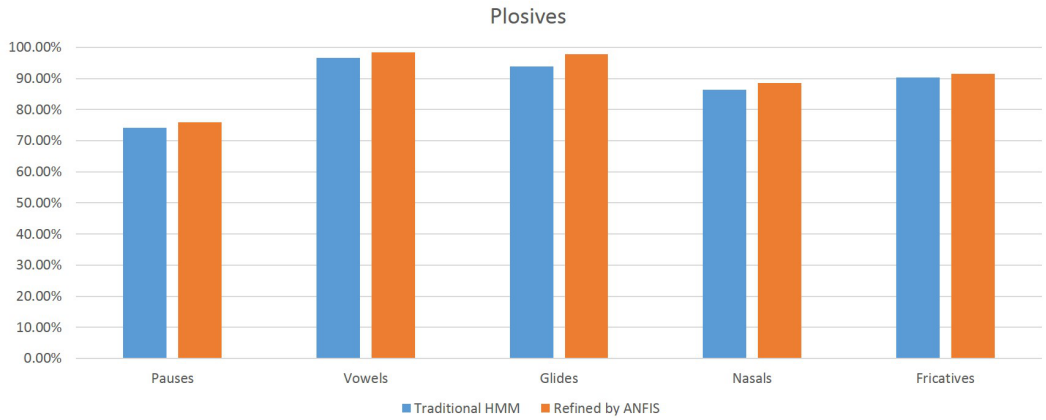


Figure 5.10: The accuracy of boundaries between Plosives and others within 20ms tolerance.

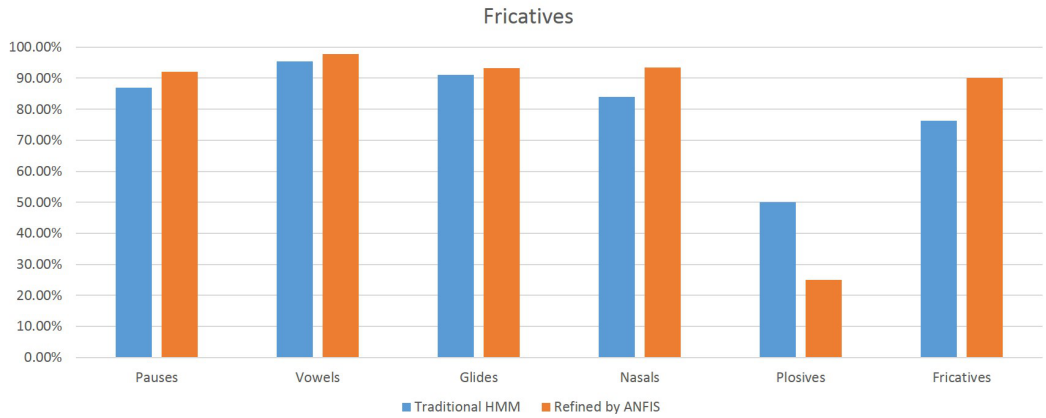


Figure 5.11: The accuracy of boundaries between Fricatives and others within 20ms tolerance.

Chapter 6

Discussion

Study the process of human segmentation of speech is a practical way to improve automatic speech segmentation. ANFIS has the ability to learn the relationship between the effective inputs and output. The accuracy rate of final results as shown in Fig. 6.1. We can see the ANFIS used in our method significantly improved the accuracy of forced alignment within HMM framework, which give the evidence for our idea.

What's more, in our proposed method, the Fuzzy Inference system improved 5% accuracy compared with the HMM-based forced alignment within 20ms tolerance. Compared with the handmade fuzzy-logic rules used in the-state-of-the-art research [18], which only has 2% improvement. And also for the statistical correction procedures used in the outstanding work to get an accuracy of 89.98% [15], we obtained 92.08% within 20ms, it is better to use Adaptive Neuro-Fuzzy Inference System. Another advantage of our method is that ANFIS is easy to be built and trained, and can be quickly implemented on other speech databases.

There are two reasons that ANFIS has a better performance compared with the fuzzy logic made by experts. Firstly, ANFIS directly generate the fuzzy rules from the training data, compare with the rules made by experts, it is more reliable. Secondly, the number of rules generated by ANFIS is far more than the rules made by experts, which means that ANFIS can obtain some more detailed rules or some implicit rules.

To analysis the results for different kinds of boundaries, the accuracy rate were calculated separately as shown in Fig. 5.6 to Fig. 5.11. For example, the Fig. 5.6 is the results of boundaries between pauses and others, including the accuracy rate for the boundaries of HMM and the refinement by ANFIS.

From these six Fig.s, we can see not all kinds of boundaries improved, after use ANFIS. The accuracy of boundaries between vowels and plosives as shown in Fig. 5.7, decreased from 96.99% to 94.65% within 20ms tolerance. This phenomenon also occurs in the following types of boundaries: boundary between glides and plosives, boundary between nasals and plosives, boundary between plosives and plosives and boundary between fricatives and plosives. Compare these kinds of boundaries with others, we found that the samples of these types of boundaries are far less than others in the training data. This factor may lead to the final results of reduced accuracy.

The second phenomenon ,that we can conclude from the results, is the accuracy of

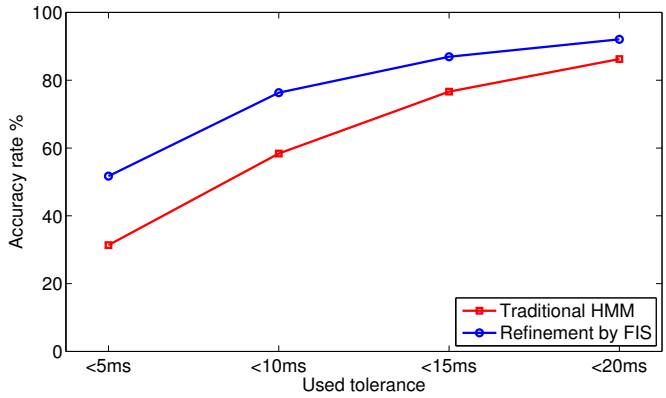


Figure 6.1: The comparison between manual segmentation and automatic speech segmentation.

boundaries between the same type of phonemes is far less than others. Firstly, these kinds of boundaries have a less number of samples in training data. Otherwise, to determine the boundaries between the same type of phonemes is also difficult in manual speech segmentation.

Although the proposed method can obtain more precise boundaries compared with traditional HMM-based forced alignment, there are some drawbacks for the proposed method.

1. This proposed method is still a top-down method for phonetic segmentation, the bottom-up information of speech signal should be taken seriously.

2. Although manual labelled time mark is always considered the right answer for speech segmentation, there should be a subjective evaluation method for the proposed method. For example, speech synthesis is carried out using the segments obtained by the proposed method, and then the naturalness of synthesized speech will be checked through listening tests.

Chapter 7

Conclusion and future work

7.1 Summary

In this thesis, a method based on ANFIS is proposed to obtain more precise phoneme boundaries close to hand-made boundaries for automatic speech segmentation, in order to supply automatic segmented speech corpus for highly natural speech synthesis or some other applications with the requirement for highly accurate and reliable speech segmentation. The results of experiments for segmentation indicates the effectiveness of the ANFIS for the purpose of this research. Moreover, our method is easy to be built and applied to other databases. For the future work, how to make the system effective and easy to be built can be the next topic.

Some small questions are investigated in the experiments, including the use of PLP and MFCC, the frame rate used for the acoustic feature extraction. The results show that there is no much difference between PLP and MFCC for the training of HMMs. In addition, the frame rate influence the accuracy for training, but not the smaller, the better. As same as the acoustic features used for the training of ANFIS, parameters need to be optimized and balanced.

In the discussion, we can see different kinds of boundaries have different accuracy for both HMM and ANFIS. As we all know, different kinds of boundaries always different in the degree of difficult in manual speech segmentation. For the reason that ANFIS is to learn the knowledge of human beings from training data. If there is no clear rules used for the segmentation, ANFIS could not performance well. It also means, the limitation of this method is the knowledge in training data. For some special situation, the accuracy decreased after refine by ANFIS. It may caused by the number of samples used for training. For some kinds of boundaries, there is too less samples in TIMIT database.

7.2 Future work

Although, we have improved the accuracy of HMM-based speech segmentation on TIMIT database. Considering the proposed method may be provided for different databases, apply this method in other databases should be the next work. Another important future

work is to combine it with other boundary detection technology to obtain more reliable boundaries. Also, the features used for training and the parameters used for the extraction of features can be continue optimized.

Consider the influence from the number of samples in this study. We can investigated the relationship between the number of samples used for training and accuracy later. Besides, some kinds of boundaries are rather ill defined in speech corpus. How to deal with these kinds of boundaries should be considered.

For the reason that we do not have enough data used for training. In this study, only 54 ANFIS were trained for each phoneme. Construct the ANFIS for each kind of boundary can be the next topic, if we have enough data used for training.

As we all know, the ANFIS extract the rules for the refinement of time marks of boundaries. But we are not clear know these rules. How to transform these rules into human knowledge and compare these with the knowledge used by experts, in order to establish complete knowledge system for speech segmentation should be considered.

7.3 Contribution

This study proposed a new method to improve the HMM-based speech segmentation. It proved the effectiveness for the use of ANFIS in automatic speech segmentation. Further more, by extract the human-liked rules from the corpus, it offers a new way to complete the knowledge for speech segmentation.

Acknowledgements

With the completion of this thesis, my master course at School of Information Science in JAIST will be close to the end. I would like to express my sincere gratitude to all those who have helped me for my study and research during the past year.

Bibliography

[1] I. Mporas, T. Ganchev, and N. Fakotakis. Speech segmentation using regression fusion of boundary predictions. Computer Speech & Language, 24(2):273–288, 2010.

[2] Speech corpus. http://en.wikipedia.org/wiki/Speech corpus.

[3] D.T. Toledano, L.A.H. Gomez, and L.V. Grande, Automatic phonetic segmentation, IEEE Trans. Speech and Audio Proc., 11, pp. 617-625, 2003.

[4] A. Acero, 1995. The role of phoneticians in speech technology. In: Bloothooft, G., Hazan, V., Huber, D., Llisterri, J. (Eds.), European Studies in Phonetics and Speech Communication. OTS Publications

[5] K. Kvale, Segmentation and Labeling of Speech, Ph.D. Dissertation, The Norwegian Institute of Technology, 1993.

[6] Aversano, G., Esposito, A., Esposito, A., Marinaro, M., 2001. A new text-independent method for phoneme segmentation. In: Proceedings of 44th IEEE Midwest Symposium on Circuits and Systems, vol. 2, pp. 516?519

[7] Adami, A.G., Hermansky, H., 2003. Segmentation of speech for speaker and language recognition. In: Proceedings of Eighth European Conference on Speech Communication and Technology (EUROSPEECH 2003), pp. 841?844.

[8] Paulo, S., Oliveira, L.C., 2003. DTW-based phonetic alignment using multiple acoustic features. In: Proceedings of eighth European Conference on Speech Communication and Technology (EUROSPEECH 2003), pp. 309?312.

[9] Keshet, J., Shalev-Shwartz, S., Singer, Y., Chazan, D., 2007. A large margin algorithm for speech-to-phoneme and music-to-score alignment. IEEE Transactions on Audio, Speech, and Language Processing 15 (8), 2373?2382.

[10] F. Brugnara, and D. Falavigna,and M. Omologo, "Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models," *Speech Commun.*, vol. 12, pp. 357-370, Aug 1993.

[11] J. P. Hosom, "Automatic Time Alignment of Phonemes Using Acoustic-phonetic Information," *PhD thesis, Oregon Graduate Institute of Science and Technology.*, 2000.

[12] D. T. Toledano, "Neural network boundary refining for automatic speech segmentation," *in Proc. ICASSP.*, pp. 3438-3441, 2000.

[13] K. S. Lee, "MLP-based phone boundary refining for a TTS database," *in IEEE Trans.*, vol. 14, no. 3, pp. 981-989 May 2006.

[14] H. Y. Lo and H. M. Wang., "Phonetic boundary refinement using support vector machine," *in Proceedings of ICASSP.*, pp. 933-936 2007.

[15] Yuan, Jiahong and Ryant, Neville and Liberman, Mark and Stolcke, Andreas and Mitra, Vikramjit and Wang, Wen, "Automatic phonetic segmentation using boundary models," *INTERSPEECH.*, vol. 12, pp. 2306–2310, 2013.

[16] Andreas Stolcke, Neville Ryant, Vikramjit Mitra, Wen Wang, and Mark Liberman, "HIGHLY ACCURATE PHONETIC SEGMENTATION USING BOUNDARY CORRECTION MODELS AND SYSTEM FUSION ," *ICASSP.*, May 2014.

[17] S. Raptis and G. Carayannis, Fuzzy Logic for RuleBased Formant Speech Synthesis, Proc. EUROSPEECH 1997, pp. 1599-1602.

[18] D. T. Toledano, M. A. Rodrguez Crespo, J. G. Escalada Sardina "Trying to Mimic Human Segmentation of Speech Using HMM and Fuzzy Logic Post-correction Rules," *Proceedings of third ESCA/COSCOSDA International Workshop on Speech Synthesis.*, Nov. 1998.

[19] D. T. Toledano, Neural network boundary refining for automatic speech segmentation, presented at the Proceedings of the International Conference on Acoustics Speech and Signal Processing 2000, Istanbul, Turkey, June 2000.

[20] J.S. Garofolo, TIMIT Acoustic-Phonetic Continuous Speech Corpus(LDC93S1), Linguistic Data Consortium, 1993.

[21] TIMIT. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp/catalogId=LDC93S1.

[22] J.P. Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech Communication.*, pp. 352-368, 2009.

[23] H. Hermansky and N. Morgan, RASTA processing of speech, IEEE Trans. Speech Audio Proc., vol.2, pp.578-589, 1994.

[24] S. Young, G. Evermann, M. Gales, T. Hain, X. Liu, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. The HTK Book (for HTK Version 3.4), Cambridge, U.K.: Cambrige Univ., March, 2009.

[25] V.W. Zue, and S. Seneff, Transcription and alignment of the TIMIT database, in H. Fujisaki (ed.), Recent Research Towards Advances Man-Machine Interface, pp. 515-525, 1996.

[26] A. Esposito, E.C. Ezin and C.A. Reyes-Garcia, Designing a fast Neuro-fuzzy system for speech noise cancellation. Lecture Notes Comput. Sci. 2000.

[27] JE Munoz-Exposito, S. Garcia-Galan, N. Ruiz-Reyes, P. Vera-Candeas, Adaptive network-based fuzzy inference system vs. other classification algorithms for warped LPC-based speech/music discrimination.Eng. Appl. Artif. Intell.2007;20:783-793, 2007.

[28] S. Cox, R. Brady, and P. Jackson, Techniques for accurate automatic annotation of speech waveforms, in Proceedings of the International Conference on Spoken Language Processing, vol. V, Sydney, NSW, Australia, 1998, pp. 1947-1950.

[29] J. Shing and R. Jang, ANFIS: Adaptive Network Base Fuzzy Inference System, in IEEE Trans. Syst. Man, Cybern. , 1993, vol. 23, pp. 665-685, 1993.

[30] T. Takagi and M. Sugeno, Derivation of fuzzy control rules from human operators control actions, in Proc. IFAC Symp. Fuzzy Inform, Knowledge Representation and Decision Analysis, pp. 55-60, July 1983.

[31] MATLAB version 7.14. Natick, Massachusetts: The MathWorks Inc., 2012.

Publications

[1] Liang Dong, Reda Elbarougy and Masato Akagi, HMM-based phonetic segmentation using Adaptive Neuro Fuzzy Inference system, Proceedings of the Autumn Meeting of the ASJ, September 2014.
[2] Liang Dong, Reda Elbarougy and Masato Akagi, Accurate phoneme segmentation method using combination of HMM and Fuzzy Inference system, Proceedings of IEICE Speech Meeting, SP2014-57, June 2014.